Self-supervised representation learning from 12-lead ECG data

Temesgen Mehari [a,b,1], Nils Strodthoff [b,1,*]

[a] *Physikalisch Technische Bundesanstalt, Berlin, Germany*
[b] *Fraunhofer Heinrich Hertz Institute, Berlin, Germany*

ABSTRACT

Clinical 12-lead electrocardiography (ECG) is one of the most widely encountered kinds of biosignals. Despite the increased availability of public ECG datasets, label scarcity remains a central challenge in the field. Self-supervised learning represents a promising way to alleviate this issue. This would allow to train more powerful models given the same amount of labeled data and to incorporate or improve predictions about rare diseases, for which training datasets are inherently limited. In this work, we put forward the first comprehensive assessment of self-supervised representation learning from clinical 12-lead ECG data. To this end, we adapt state-of-the-art self-supervised methods based on instance discrimination and latent forecasting to the ECG domain. In a first step, we learn contrastive representations and evaluate their quality based on linear evaluation performance on a recently established, comprehensive, clinical ECG classification task. In a second step, we analyze the impact of self-supervised pretraining on finetuned ECG classifiers as compared to purely supervised performance. For the best-performing method, an adaptation of contrastive predictive coding, we find a linear evaluation performance only 0.5% below supervised performance. For the finetuned models, we find improvements in downstream performance of roughly 1% compared to supervised performance, label efficiency, as well as robustness against physiological noise. This work clearly establishes the feasibility of extracting discriminative representations from ECG data via self-supervised learning and the numerous advantages when finetuning such representations on downstream tasks as compared to purely supervised training. As first comprehensive assessment of its kind in the ECG domain carried out exclusively on publicly available datasets, we hope to establish a first step towards reproducible progress in the rapidly evolving field of representation learning for biosignals.

1. Introduction

The availability of datasets with high-quality labels is an omnipresent challenge in machine learning in general, but especially in the health domain, where the labeling process is particularly expensive and clinical ground truth is in many cases hard to define. However, the amount of unlabeled data often exceeds the amount of labeled data by several orders of magnitude, which represents a strong case for (self-supervised) representation learning from unlabeled data. During the past few years, self-supervised learning has made enormous advances in different domains ranging from natural language processing [1] over speech [2] to computer vision [3]. Self-supervised learning could be one component towards addressing the problem of data scarcity. It could help to train more accurate and potentially also more robust models given the same amount of labeled data, which is a desirable prospect for

any application field. Of particular importance for the medical domain are improvements in label efficiency, which could allow to train models on more finegrained and consequently less populated label hierarchies, or to include rare diseases that were out of reach with conventional training methods.

In this work, we investigate self-supervised representation learning in the context of clinical electrocardiography (ECG) data. The ECG is a non-invasive method that allows to assess the general cardiac condition of a patient. It is therefore an important tool for the first-in-line examination for the diagnosis of cardiovascular diseases, which rank among the diseases of highest mortality [4]. In particular, the (short) 12-lead ECG, which we focus on in this work, is the most commonly used type of ECG with a very broad clinical applicability ranging from primary care centers to intensive care units. Even though the technology underlying the ECG is by now more than 100 years old and it is an

extremely common procedure, which is ordered or provided during 5% of the office visits in the US [5], its interpretation is still performed mostly manually with only limited algorithmic support. Here, it is important to recognize that ECG interpretation is in some cases even challenging for cardiologists [6].

There are deep-learning-based ECG interpretation algorithms with exceptionally high performance [7,8] that have been trained on large closed-source datasets. The sizes of publicly available datasets are smaller by several orders of magnitude, which provides the motivation to investigate if and how far self-supervised learning techniques can improve the performance of algorithms trained on these datasets. In addition, the question of label quality remains challenging even for the above large-scale datasets. Here, it is important to stress that even though self-supervised methods have been applied successfully in computer vision and speech, ECG records are timeseries (rather than a one-dimensional image) and multivariate data (unlike speech) with considerably different properties than speech. This means that the degree to which self-supervised methods work in this domain is not clear from the onset and deserves a thorough study. Even though the underlying methods are based on methods developed for other application domains, it requires subtle adaptations, such as a careful choice of augmentation transformations or adaptations of the model architecture and training procedure, to make them actually work in the context of ECG data. And finally, just as in the case of supervised learning [9], measurable progress in the field of representation learning for ECG data requires benchmarking based on clearly defined evaluation criteria, in the ideal case with open software on publicly accessible datasets. With this work, we hope to establish a first step in this direction.

Going beyond a mere benchmarking of representation learning algorithms for the ECG domain, the real benefits lie in the potential benefits of self-supervised pretraining for finetuned downstream classifiers. This includes aspects such as including improved data efficiency, improved quantitative performance, or improved robustness in a general sense as compared to model trained in a purely supervised fashion. In our experimental results, we present explicit evidence for these benefits. Putting these results into perspective, demonstrating significant improvement through self-supervised pretraining in terms of downstream performance should not be taken for granted as the effects often remain small [10]. Also improved robustness properties from self-supervised have rarely been demonstrated explicitly in other domains not to mention the domain of ECG data.

Our key achievements can be summarized as follows:

- We present the first comprehensive assessment of self-supervised representation learning for 12-lead ECG data to foster measurable progress in the subfield of representation learning for biosignals.
- We adapt and directly compare instance-based self-supervised methods (Simple Framework for Contrastive Learning of Visual Representations *(SimCLR)*, Bootstrap Your Own Latent *(BYOL)*, Swapping Assignments between multiple Views *(SWaV)*) and contrastive, latent forecasting methods (contrastive predictive coding *(CPC)*) and find compelling evidence for the feasibility of learning useful representation from ECG data through self-supervised learning.
- We propose and evaluate several modifications in the CPC architecture and training procedure that lead to considerable performance improvements.
- We evaluate different quality aspects of downstream classifiers finetuned from self-supervised models compared to training from scratch and find evidence for improved quantitative performance given the same downstream training set, improved label efficiency and improved robustness through self-supervised pretraining.

2. Materials and methods

2.1. ECG analysis using deep learning

By now, the analysis of ECG data has developed into a very popular application domain for deep learning methods. As we are solely focusing on deep learning methods in this work, a brief discussion of state-of-the-art methods in the field is appropriate. For a detailed comparison of different approaches, we refer the reader to recent reviews [11,12]. Until very recently, the merits of newly proposed methods in the field have been very difficult to assess due to a lack of appropriate, large, publicly available ECG datasets for training and evaluation of algorithms as well as due to a lack of clearly defined benchmarking criteria. The first issue was resolved very recently with the publication of several large clinical ECG datasets, see Section 2.3 for details. Relating to the second issue, we draw on a recent benchmarking study [9] on the *PTB-XL* dataset [13], the very dataset also used as downstream dataset in this study, where a number of different classification algorithms was assessed on different clinically relevant ECG classification tasks. The overall best-performing methods in this study turned out to be modern convolutional network architectures, namely resnet- or inception-based architectures. This is in line with the network architectures used in the literature [7,8,14] for which the authors report excellent performance on datasets with restricted access that are up to several orders of magnitude larger than the currently available public datasets.

2.2. Self-supervised representation learning for ECG data

Contrastive methods in computer vision have witnessed tremendous advances in the past few months [3,15–18], which significantly improved the linear evaluation performance on ImageNet and demonstrated the usefulness of the learned features for other computer vision tasks. These methods can be adapted straightforwardly to learning representations from a large number of relatively short time series segments if one interprets the time series record as a one-dimensional multichannel image and adapts transformations appropriate for time series. The predominantly used approaches rely on instance discrimination as pretraining task and will be discussed in Section 2.2.1 in detail. A second domain, where self-supervised methods for non-discrete data have been implemented successfully is the domain of representation learning speech, where predictive coding methods [2,19] have been applied to conventional acoustic features [19–21] but also to raw waveform data [2,22,23]. The best-performing methods in this field rely on latent forecasting tasks and are discussed in Section 2.2.2.

2.2.1. Instance discrimination (SimCLR/BYOL/SwAV)

Current state-of-the-art contrastive methods from computer vision aim to learn representations based on multiple views on the same instance, see Fig. 1. These are created by applying stochastic transformations to the input data. This idea is implemented in the most straightforward way in *SimCLR* [3], where a noise contrastive loss is used to attract two (positive) copies originating from the same original instance and to repel instances from all other (negative) instances in the batch, an approach which typically relies on training with large batch sizes, which is less problematic in our case due to the reduced dimensionality of time series data as compared to image data. *BYOL* [17] does not explicitly rely on contrasting against negative samples in the same batch, but uses a moving average of the model itself and reported slight improvements over *SimCLR* in the image domain. Finally, *SwAV* [18] relies on contrasting cluster assignments rather than individual instances and once again improved the linear evaluation scores on ImageNet. In our case, we build on the implementations of all three frameworks in *PyTorch Lightning Bolts* [24].

As model architecture, we use the convolutional neural networks of the *xresnet1d*-family, one-dimensional adaptations of the popular *xresnet*-architecture [25] from computer vision, which showed very good
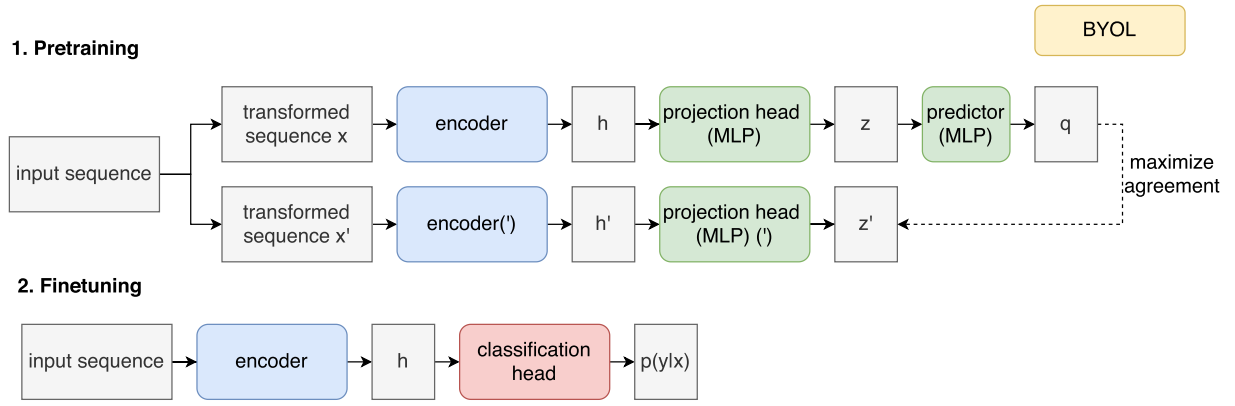
1. Pretraining

2. Finetuning

Fig. 1. Schematic illustration of contrastive methods based on instance-discrimination for the case of BYOL.

perform in a very recent ECG classification benchmarking study [9] that was carried out on the *PTB-XL* dataset, see Section 2.3, using the same evaluation scheme that is supposed to be used in this work, see Section 2.4. We base our experiments on a *xresnet1d50*, whose performance is compatible with that of the best-performing *xresnet1d101* from Ref. [9] within error bars. However, it is more parameter-efficient and showed a slightly superior performance for linear evaluation and finetuning. At this point, we stress again that even though we use an architecture that achieves state-of-the-art performance on *PTB-XL*, the main point of our study lies in the demonstration of relative improvements compared to supervised performance.

The transformations used to generate two semantically equivalent views of a given original record lie at the heart of the recent success of contrastive methods in computer vision. As demonstrated in Ref. [3], the quality of the learned representations depends crucially on the choice and proper combination of transformations. We therefore evaluated a number of transformations inspired by effective transformations in computer vision and transformations specific for time series, see Appendix A for a detailed description. Finally, we also evaluate representations obtained by using only prototypical physiological noise during pretraining.

2.2.2. Latent forecasting (CPC)

Contrastive Predictive Coding (*CPC*) [2] is also a contrastive approach, which, in contradistinction to the approaches described above, explicitly makes use of the sequential ordering of the data. The idea is to encode the input sequence by means of an encoder with strided convolutions or fully connected layers and train a model to predict the latent representation of the sequence a fixed number of steps in the future given the encoded representation of the sequence in the past again using a noise contrastive estimation approach, see Fig. 2 for a graphical representation. As we work with data at 100 Hz, which is sampled rather coarsely compared the typical sampling rates of 10 kHz in the audio domain, there is no need to drastically downsample the

signal by means of strided convolutions. Instead, we use a fully connected encoder, in our case composed of four layers with 512 filters with batch normalization, as it was also done in self-supervised representation learning from classical audio features [19–21]. We predict 12 timesteps or equivalently to 0.12s into the future and work with 128 false negatives that are drawn from the same sequence as the original record. For the prediction task, we use a LSTM model [26] with 2 layers and 512 hidden units. We propose and evaluate an enhanced version of the *CPC* architecture, with an additional hidden layer and non-linearity before the linear output layer of the LSTM. This modification was inspired by the additional multilayer perceptron in *SimCLR*, which was one of the key components that lead to superior performance compared to previously used self-supervised approaches in computer vision.

When finetuning a classification model, we apply a concat-pooling layer [27], which concatenates the maximum of all LSTM outputs, the mean of all LSTM outputs, and the LSTM output corresponding to the final step, and a fully connected classification head with a single hidden layer with 512 units including batch normalization and dropout for regularization. To assess the linear evaluation performance, we use a single fully connected layer on top of the concat-pooling layer. The effect of the different modifications compared to standard *CPC* implementations and finetuning schedules are investigated in detail in Appendix C.

2.2.3. Self-supervised representation learning for physiological time series data

Self-supervised methods have also been used for representation learning from biomedical sequence data, including, most prominently, ECG [28–31] and electroencephalography (EEG) [28,29,32,33] data. With the exception of [31], none of the existing works considered the case of representation learning from clinical 12-lead ECGs, the clinically most widely encountered type of ECG measurement. The authors of [31] also consider BYOL and *SimCLR* for pretraining but used a very shallow network architecture with only five layers. We believe that it is necessary to use larger models, which reach state-of-the-art performance on
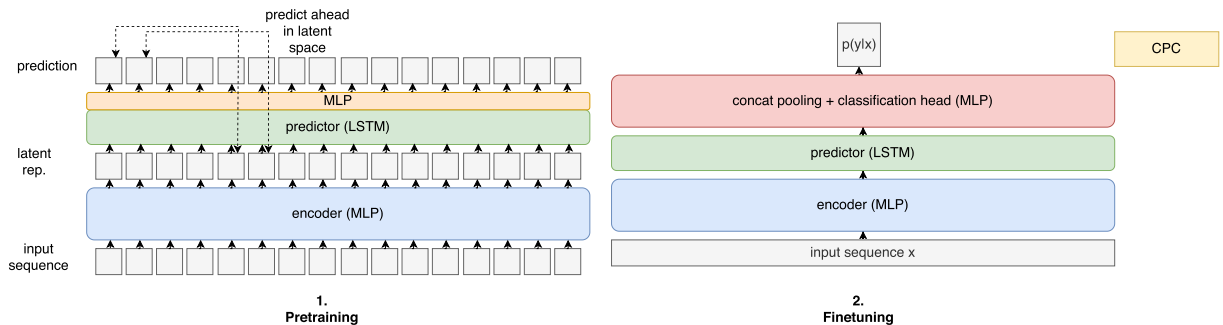

Fig. 2. Schematic illustration of pretraining and finetuning procedure for contrastive predictive coding (*CPC*).

large, comprehensive ECG datasets such as *PTB-XL* and which consequently allow to learn richer representations, along with pretraining on larger datasets. In particular, it is not clear if pretraining advances in the small model regime carry over to larger models with competitive supervised performance. Also, from the methodological point of view [31], deviates significantly from our approach in the sense that they aim to learn a lead-independent, universal single-lead encoder, as opposed to a joint 12-lead encoder in our case. The former is not directly applicable to downstream 12-lead ECG tasks. In addition, they propose new contrastive methods that can use 12-lead ECG data during pretraining but differ from our methods in that they are not expedient for downstream 12-lead ECG tasks. This is because the proposed models do not process 12-lead data directly but exploit the fact that different leads from the same patient during pretraining can be considered as positive pairs. From the methodological point of view [29], is also close to our contrastive approach but their experimental results were limited to a small 2-lead dataset with less than 50 records. Without access to the original implementation, it is impossible to assess if their proposed approach would be competitive on 12-lead data and on large (pretraining) datasets, where self-supervised methods reveal their full potential. Earlier approaches such as [28] trained representations from 2-lead ECGs using skip-gram models. Finally [30], use transformation recognition as a pretext task and proposed a framework specific to representation learning from single-lead ECGs.

2.3. ECG datasets

We use a collection of three datasets for pretraining henceforth referred to as *All*, namely *CinC2020* [34], *Ribeiro* [8] and *Chapman* [35], which constitutes a collection of the largest publicly available 12-lead ECG datasets with in total 54,566 records. It is worth mentioning that *CinC2020*, the training dataset used for the Computing in Cardiology Challenge 2020, is by itself a compilation of five different datasets. In particular, it includes the *PTB-XL* dataset [13,36] that we also use for evaluation in this study. At the most finegrained level, which is used here, the *PTB-XL* dataset comes with 71 labels and the evaluation task is framed as a multi-label classification task. It is worthwhile stressing that these labels cover a wide variety of diagnostic, form and rhythm statements and can be used for a comprehensive evaluation of ECG analysis algorithms. The 44 diagnostic statements can be categorized in terms of five super classes (normal/conduction disturbance/myocardial infarction/hypertrophy/ST-T change), the 19 form statements relate to mostly morphological changes in specific ECG segments such as an abnormal QRS complex, and the 12 rhythm statements comprise statements characterizing normal cardiac rhythms as well as arrhythmia. The dataset is organized into ten stratified, label-balanced folds, where the first eight are used as training set, the ninth is used as validation set and the tenth fold serves as test set [13]. All datasets are summarized in Table 1.

2.4. Training and evaluation protocol

We restrict ourselves to ECG data at a sampling rate of 100 Hz in all cases. We pretrain CPC models on input sequences with a length of 10 s,

Table 1
Dataset summary: For pretraining we use *All* (*CinC2020* and *Chapman* and *Ribeiro*) or *PTB-XL*. We evaluate on *PTB-XL*. Note that *PTB-XL* is a subset of *CinC2020*.

dataset	# samples	# patients
Pretraining: *All*	54,566	unknown
- *CinC2020* [34]	43,093	unknown
- *Chapman* [35]	10,646	10,646
- *Ribeiro* [8]	827	827
Evaluation	21,837	18,885
- *PTB-XL* [13]	21,837	18,885

all other models (including finetuned CPC models) are trained on input sequences with length of 2.5 s. During training, subsequences are randomly cropped from the input record. During test time while finetuning, we use test-time-augmentation and crop all sequences to a length 2.5 s (using a stride of 1.25 s) from the original record and take the mean of their respective output probabilities as final prediction, a method which considerably improved the model performance by approximately 0.01 in macro AUC as compared to a naive evaluation [9]. Both for finetuning and pretraining, we use the AdamW optimizer [37] with a weight decay of 0.001. During pretraining, we optimize the InfoNCE loss [2] using a constant learning rate schedule for CPC and the respective contrastive losses with a cosine learning rate schedule for *SimCLR*, *BYOL*, and *SwAV* as described in the original publications. During finetuning, we optimize binary crossentropy with a constant learning rate schedule as appropriate for a multi-label classification task and evaluate the model performance based on macro AUC as in Ref. [9], computed from the 71 labels on the most finegrained level in *PTB-XL* [13]. We perform model selection on the validation and select the model with the lowest validation loss during pretraining and highest macro AUC during finetuning. We report the respective test set score of the selected best model. The source code to reproduce all our experiments is publicly available [38].

As conventionally done in self-supervised representation learning studies, we use two different evaluation procedures, *linear evaluation* and *finetuning*. The *linear evaluation* protocol aims to assess the quality of the learned representations through the linear separability of the learned representations. To this end, we replace the classification head by a single linear layer and freeze all other layers as well as batch normalization statistics. Within the *finetuning* protocol, we investigate the usefulness of these representations for downstream tasks, where we unfreeze the classification head as well as all layers of the pretrained model. For *CPC*, we found it beneficial to follow a two-step approach during finetuning: In a first step, we finetune just the classification head for 50 epochs while keeping the remaining pretrained weights fixed but still updating batch normalization statistics. We perform model selection using validation set scores and then finetune the entire model for 20 epochs at a reduced learning rate using discriminative, i.e. layer-dependent learning rates to mitigate the danger of overwriting the information captured during pretraining, where we typically divide models into head, body and stem/encoder and reduce the learning rate by a factor of 10 compared to the respective previous layer group. Also in this case, we select the final model based on validation set scores. In all other cases, we train models for 100 epochs using a constant learning rate.

3. Experiments

The performance of self-supervised contrastive methods based on instance discrimination crucially depends on the choice of transformations that are used to create two semantically equivalent copies of the original input sequence. To determine appropriate transformations, we carried out an experiment, where we pretrained a model using *SimCLR* and different combinations of transformations and evaluated the quality of the learned representations based on linear evaluation performance on *PTB-XL*, see Appendix B in the supplementary material for details. The results clearly identify *time out (TO)* in combination with *random resized crop (RRC)* as most effective transformation pair, see Appendix A in the supplementary material for a description of all transformations under consideration. In addition, we consider physiological noise transformations that were designed to mimic typical physiological noise that might occur during ECG measurements, namely *baseline wander, powerline noise, electromyographic noise* and *baseline shift*. In a second step, we used a similar protocol to compare the three different contrastive learning frameworks *SimCLR*, *BYOL* and *SwAV* now using the predetermined set of transformations. Whereas *SimCLR* reached clearly the best linear evaluation performance, finetuning from

a *BYOL* representation lead to a superior downstream performance after finetuning, which is why we consider both methods in the following sections.

3.1. Self-supervised pretraining learns meaningful representations from ECG data

We start by discussing the linear evaluation performance in Table 2, which should be set in perspective to the supervised performance achieved on *PTB-XL*. The best published result for this task on the same dataset with identical splits using purely supervised training was 0.925 (07), also using a *xresnet1d*-model [9]. Our supervised results remains compatible within error bars with this baseline result. The architecture used for CPC pretraining (denoted by 4FC+*2LSTM*+2FC) was not investigated in previous studies [9] and shows the strongest supervised performance reported on *PTB-XL* thus far.

The linear evaluation performances in Table 2 show that the pre-trained representations are highly relevant for downstream classification tasks. Most notably, the linear evaluation performance of the *CPC* model only shows a performance gap of 0.5% compared to the same model architecture trained in a supervised manner and already exceeds the best result previously obtained using supervised training on the same dataset [9]. The contrastive methods from computer vision show a slightly weaker performance, but still the best linear evaluation performance reaches 95.5% of the respective supervised performance. The main point we are trying to convey here is how far one can push the linear evaluation performance not only in relative comparison to supervised performance, but also on an absolute scale. Whereas the former can also be demonstrated with simpler model architectures, the latter requires a certain model complexity. Based on these results, it is justified to claim that self-supervised representation learning is very effective in the ECG domain. To demonstrate the impact of the dataset size, we also report results for pretraining *CPC* just on *PTB-XL* i.e. using only 40% of the original training dataset. As expected, increasing the size of the training dataset leads to improvements in the linear evaluation performance. Again, the reader is referred to Appendix C in the supplementary material for details on the impact on the modifications in the original *CPC* architecture and finetuning schedule.

3.2. Self-supervised pretraining improves downstream performance

In this section, we investigate whether finetuning from self-supervised representations can potentially lead to improvements in downstream performance as compared to purely supervised training. The results are compiled in Table 2 and summarized graphically in Fig. 3. As before, *SimCLR* reaches the best linear evaluation performance whereas it is slightly outperformed by *BYOL* in terms of downstream
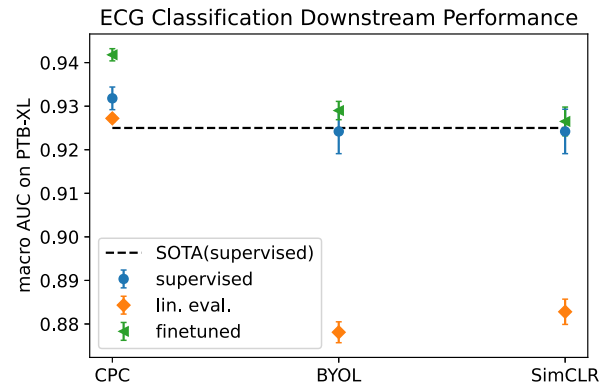

Fig. 3. Comparison of different contrastive learning frameworks in terms of downstream performance comparing the three self-supervised learning frameworks *CPC*, *BYOL* and *SimCLR*. The previous supervised state-of-the-art result from Ref. [9] is represented by a dashed line.

performance. The considerably better linear evaluation performance of the *CPC* model as compared to *BYOL* and *SimCLR* directly translates into an improved downstream performance. Interestingly, the *SimCLR* and *BYOL* finetuned model performance when using physiological noise during training almost reaches the results from using (RRC,TO)-transformations, while a sizable performance gap exists between them in terms of linear evaluation performance. In order to test in how far the overlap between using *PTB-XL* for pretraining and for evaluation leads to an overestimation of the positive effects of pretraining on unseen data, we also investigate the performance of a model pretrained on *CinC2020 w/o PTB-XL*, i.e. *CinC2020* with records from *PTB-XL* excluded, which is comparable in size to *PTB-XL*. On the one hand, the linear evaluation performance after pretraining on *CinC2020 w/o PTB-XL* is lower and does not overlap with the result from pretraining on *PTB-XL*, which supports the initial hypothesis. On the other hand, the seemingly superior representation from pretraining on *PTB-XL* do not translate into a stronger classification performance after finetuning. In fact, finetuning from pretraining on *CinC2020 w/o PTB-XL* even leads to a higher point estimate whereas both results remain consistent within error bars. Hence, these somewhat inconclusive results do not allow any decisive statements about the initial hypothesis. In particular, it remains difficult to disentangle the effects of overlap of samples in pretraining and finetuning from differences in the distribution of the pretraining datasets, which are rather pronounced as shown in Fig. E1.

In all cases, the results from finetuning pretrained models improve over the corresponding supervised results (by 1.0% for *CPC*, by 0.2% for *SimCLR*, and by 0.5% for *BYOL*). Furthermore, it is noticeable that already after the first finetuning step, where just the batch norm statistics and the classification head were adjusted, the *CPC* model reaches performance values around 0.931, i.e. already roughly matches supervised performance. These results provide a clear case for self-supervised learning in the ECG domain.

At this point we find it appropriate to comment on different sources of uncertainty impacting our results. In Table 2, we report uncertainties related to the inherent randomness of the training process, which we assess via multiple training runs. In addition to this systematic error, there is also an uncertainty in the final scores due to the finiteness and the particular sample distribution of the test set. As in Ref. [9], we assess this error via empirical bootstrapping on the test set using 1000 bootstrap iterations and evaluate 95% confidence intervals. For finetuning after self-supervised pretraining we find confidence intervals ± 0.006 (± 0.008 for linear evaluation) as opposed to ± 0.008 for training from scratch with only minor variations between different runs. This statistical error therefore represents the dominant source of uncertainty, which provides a strong argument for larger ECG evaluation datasets. As in Ref. [9], we check if the confidence interval for the difference

Table 2
Linear evaluation and finetuning performance on a comprehensive clinical downstream ECG classification task (macro AUC on the *PTB-XL* test set). As before, we report mean and standard deviation over 10 finetuning runs.

method	model	PTB-XL	
		lin. eval.	finetuned
Supervised	4FC+2LSTM+2FC	0.7110(65)	0.9318(26)
Supervised	xresnet1d50	0.7210 (158)	0.9242(51)
CPC (on All)	4FC+2LSTM+2FC	**0.9272(08)**	**0.9418 (14)**
CPC (on PTB-XL)	4FC+2LSTM+2FC	0.9246(08)	0.9395(11)
CPC (on CinC2020 w/o PTB-XL)	4FC+2LSTM+2FC	0.9192(11)	0.9401(16)
SimCLR (RRC, TO)	xresnet1d50	0.8828(29)	0.9265(33)
SimCLR physio.	xresnet1d50	0.7701(31)	0.9258(13)
BYOL (RRC,TO)	xresnet1d50	0.8781(24)	0.9290(21)
BYOL physio.	xresnet1d50	0.8295(27)	0.9260(25)

between finetuning following pretraining and training from scratch encloses only positive values, which would indicate that the performance improvement is statistically significant. We investigated the improvements for every combination of the 10 models that were finetuned from the model pretrained on *All* and 10 models trained in a supervised fashion, combining both sources of uncertainty in a single analysis. In 90% of these 100 comparisons, pretraining led to a statistically significant improvement, underscoring the positive effects of self-supervised pretraining.

In Fig. 4, we break down the macro AUC values of the best performing model by ECG statement. Per statement, we present both the performance of the supervised trained model (in color) and the improvement (in black) that occurs through self-supervised training. Moreover, we color the statements by their respective super class (for diagnostic statements), or mark them as rhythm or form statement and sort them according to the supervised performance. In terms of improvements through self-supervised pretraining, the largest average gains are observed for form and rhythm statements, see, Fig. D1 which supports the hypothesis that different pathologies tend to profit differently from self-supervised pretraining. This effect is superimposed by a different effect that is visible from Fig. 4, which indicates that the gains through pretraining show a negative correlation with the performance level before pretraining, i.e. ECG statements with low supervised performance tend to profit more from self-supervised pretraining, see Fig. D2 for an explicit demonstration.

As a final remark, one has to consider the different sizes of models under consideration. Whereas the *CPC*-model during finetuning comprises 5.8 M parameters, the *xresnet1d50* only counts 930 K parameters, which might suggest that part of the gap between *CPC* and *BYOL/SimCLR* can at least partially be attributed to a difference in model capacity. However, in preliminary experiments we saw no indications of strong performance increases with wider or deeper models. It remains to see how the performance of *SimCLR* and *BYOL* scales on larger datasets. On the contrary, for the 4FC+*2LSTM*+2FC-models, the model capacity is instrumental to reach its very high performance, see Appendix C. A
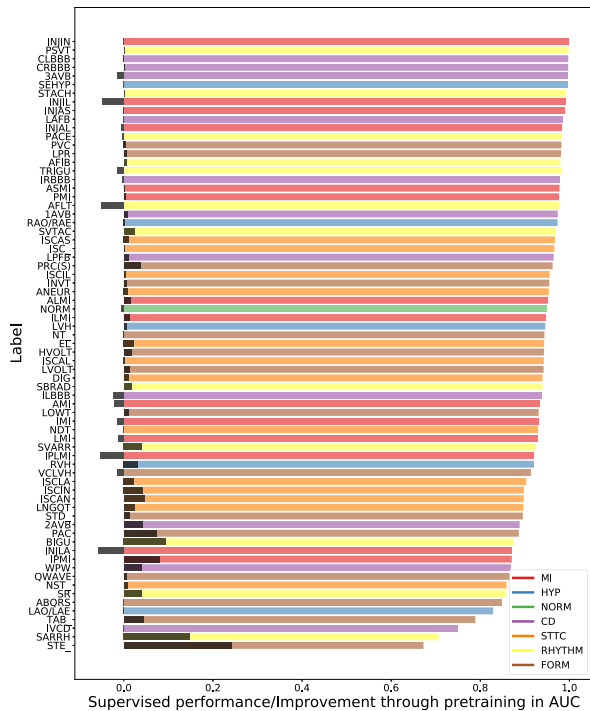
serious disadvantage of *CPC* is the sequential nature of the LSTM, which leads to slow training times. The training times are not directly comparable due to the different nature of the tasks, but give at least a hint. *CPC* models were pretrained for 200 epochs, which took approximately 6 days on a single Tesla V100 GPU. *SimCLR* and *BYOL* pretraining was performed for 2000 epochs using batch sizes of 8192 with approximate runtimes of 15 h and 13 h on a single Tesla V100 GPU, respectively. Performing $50 + 20(100)$ epochs of finetuning for the 4FC+*2LSTM*+2FC (*xresnet1d50*)-model takes approximately 25(10) minutes on the mentioned hardware.

3.3. Self-supervised pretraining improves downstream data efficiency

Another potential advantage of self-supervised pretraining lies in a potentially improved data efficiency when finetuned on a downstream task. This is a particularly relevant case for medical applications, where high-quality labels are hard to obtain. To investigate this claim in detail, we compare the performance of different pretrained models from self-supervised representations to models trained from scratch in purely supervised fashion while varying the number of training folds from the original 8 folds to a single fold. This can be read off for example from the number of training folds where the pretrained model reaches the same performance as the supervised model trained on the full dataset. In the case of *CPC*, this point is reached approximately at 5 folds or equivalently approximately 62% of the training data. The performance level of the supervised models at 4 folds is approximately reached by the pretrained model trained only on two folds i.e. 50% of the training data. For *BYOL* and *SimCLR*, the effect is still present but much less pronounced due to the closer proximity of the pretrained and the purely supervised results on the full training set. Fig. 5 nicely illustrates another advantage of self-supervised pretraining, namely the performance across different training runs is much more stable compared to purely supervised training, as visible from considerably reduced error bands, most crucially influenced by the two-step finetuning procedure in combination with discriminative learning rates, see also Appendix C in the supplementary material.

3.4. Self-supervised pretraining improves robustness of downstream classifiers

In addition to quantitative performance and data efficiency, robustness is one of the key quality criteria for machine learning models.
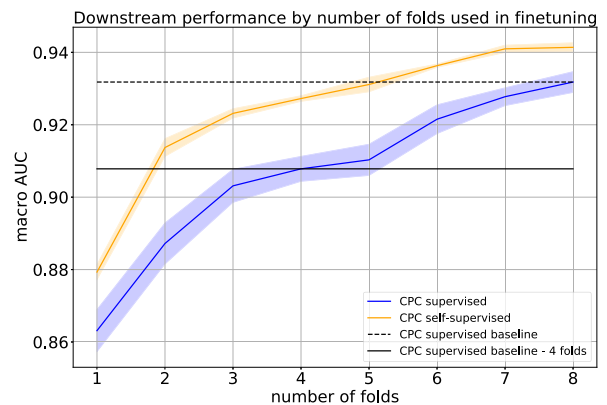

Fig. 4. Individual label AUCs for a 4FC+*2LSTM*+2FC-model from supervised training (in color, sorted in descending order) and the corresponding improvements through self-supervised training with *CPC* (in black).

Fig. 5. Finetuning downstream performance on PTB-XL dataset of a 4FC+*2LSTM*+2FC-model pretrained using *CPC* compared to its supervised counterpart. The finetuning was performed for different number of training folds, ranging from 1 to 8 folds. We used 10 runs for 8 folds as before and 3 runs for 7 folds or fewer. The plot shows the mean performance as a solid line and one standard deviation around it as a shaded band. To guide the eye, we indicate the performance of the supervised model trained on 8(4) folds performance on the full training set by a dashed(solid) line.

Here, we focus on robustness against input perturbations. It is well known that certain types of noise tend to occur in ECG data as a consequence of the measurement process and physiological interference [39,40]. In Appendix A in the supplementary material, we briefly review typical kinds of ECG noise and propose simple ways of parameterizing them. For simplicity, we just superimpose the different noise types and the original ECG waveform. We define different noise levels by adjusting the amplitudes of these noise transformations and evaluate the performance of the models from the previous sections on perturbed versions of the original test set. We also indicate signal-to-noise-ratios (SNRs) corresponding to the different noise levels, where we identify the signal with the original ECG waveform. However, one has to keep in mind that this assessment neglects the noise inherent in the original measurement, which implies that the given SNRs only upper-bound the actual values.

The goal is to investigate if pretrained models are less susceptible to physiological noise. The results in Fig. 6 reveal an interesting pattern: For the 4FC+*2LSTM*+2FC-models, the *CPC*-pretrained model shows a considerably improved robustness to noise compared to its supervised counterpart. However, both models turn out to be less robust than the considerably less complex *xresnet1d*-models. For the latter, the *BYOL*-pretrained model with physiological noise shows the strongest overall performance and also performs considerably stronger than the corresponding supervised model, which is to a certain degree an expected result as the model experienced a comparable kind of noise during pretraining. This result provides a strong argument for pretraining with domain-specific noise transformations even if it comes at the cost of a slight performance loss compared to the best-performing pretraining procedure during noiseless evaluation, see Table 2. Somewhat surprisingly, the *BYOL*-pretraining with the artificial (RRC,TO)-transformation even lead to a reduced robustness compared to the supervised *xresnet1d50*. As a final remark, the noise levels 3 and beyond are already strongly dominated by noise and correspond to situations that will rarely be encountered in real-world scenarios.

3.5. Implications of the results

Even though we discussed the implications of our findings on a technical level, we find it appropriate to also briefly comment on the broader implications of our results. Section 3.1 addresses the question how far one can push the linear evaluation performance and demonstrates that self-supervised learning can be implemented effectively in the domain of ECG data. In terms of (clinical) impact, however, it is primarily a necessary prerequisite for the following investigations: Section 3.2 shows that self-supervised pretraining improves the predictive performance compared to training from scratch. Most importantly, breaking down the improvements according to individual diagnoses shows that underperforming diagnoses when training from scratch profit most from self-supervised pretraining. This is an encouraging sign as it entails the prospects to eventually train models on even finer and hence less populated label hierarchies and/or to tackle rare diseases, for which only a very limited number of labeled samples exist in the first place. Section 3.3 demonstrates that the improvements achieved through pretraining directly translate into an improved label efficiency. This is also an encouraging result for the broader ECG research community given the growing but still compared to commercial ECG datasets small sample size available from freely accessible ECG datasets with high label quality. Finally, Section 3.4 stresses that self-supervised pretraining leads benefits such as improved robustness that go beyond quantitative performance and that are also very desirable in clinical applications.

4. Summary and conclusions

In this work, we put forward a comprehensive assessment of self-supervised representation learning on 12-lead clinical ECG data. Even though self-supervised algorithms have been applied successfully in
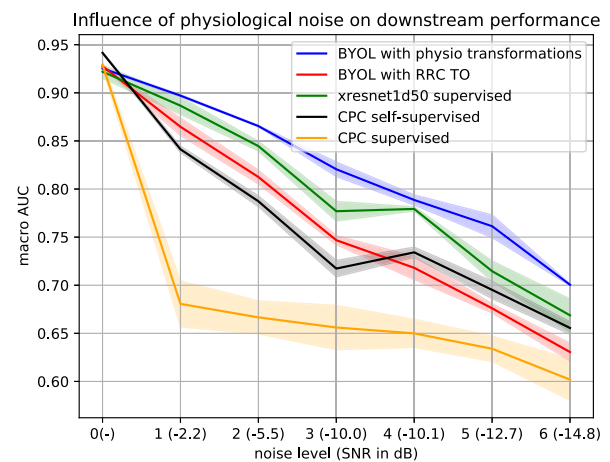

Fig. 6. Evaluating the impact of noise on pretrained and purely supervised trained classifiers. We induced typical physiological noise at different strength levels to the test set.

computer vision and speech, ECG is a different data modality, where the degree to which self-supervised learning works is not clear from the onset. Upon adapting self-supervised representation learning to the ECG domain, self-supervised learning turns out to be very effective: Self-supervised representations (via *CPC*) reach scores that only fall behind 0.5% supervised performance during linear evaluation and lead to improvements of 1.0% compared to supervised performance during finetuning, which represents a significant increase in 90% of the time within an analysis incorporating both systematic as well as statistical uncertainties. The sizable performance gap translates into an improved label efficiency, i.e. the pretrained model reaches the same performance as the supervised model but using only roughly 50–60% of the samples. We also investigate the impact of self-supervised pretraining on the robustness of the corresponding finetuned classifiers against physiological noise. We find increased robustness for most pretrained models compared to the corresponding models trained from scratch, but particularly for those that were pretrained using domain-specific noise transformations. This provides a strong case for the use of domain-specific noise transformations during pretraining.

To summarize, self-supervised learning is one path towards more robust and more label-efficient training procedures, which might alleviate the problem of label scarcity, which is particularly pressing in medical applications. In this work, we demonstrated clear advantages in terms of quantitative performance, label efficiency and robustness. It will be interesting if these carry over to further quality dimensions. We see our work as a first step towards measurable progress in the field of representation learning for 12-lead ECGs. All used datasets and the source code underlying our study are publicly available [38].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.105114.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Association for Computational Linguistics, Minnesota, Jun. 2019, pp. 4171–4186.

[2] A. van den Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, 2018 arXiv preprint 1807.03748.

[3] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020.

[4] E. Wilkins, L. Wilson, K. Wickramasinghe, P. Bhatnagar, J. Leal, R. Luengo-Fernandez, R. Burns, M. Rayner, N. Townsend, European Cardiovascular Disease Statistics 2017. plus 0.5em minus 0.4em, European Heart Network, Belgium, 2017, p. 2.

[5] CDC, National Ambulatory Medical Care Survey: 2016 National Summary Tables, Tech. Rep., Centers for Disease Control and Prevention, 2019 https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2016_namcs_web_tables.pdf.

[6] S.M. Salerno, P.C. Alguire, H.S. Waxman, Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence, Ann. Intern. Med. 138 (9) (May 2003) 751.

[7] A.H. Kashou, W.-Y. Ko, Z.I. Attia, M.S. Cohen, P.A. Friedman, P.A. Noseworthy, A comprehensive artificial intelligence–enabled electrocardiogram interpretation program, Cardiovasc. Digit. Health J. 1 (2) (Sep. 2020) 62–70.

[8] A.H. Ribeiro, M.H. Ribeiro, G.M.M. Paixão, D.M. Oliveira, P.R. Gomes, J. A. Canazart, M.P.S. Ferreira, C.R. Andersson, P.W. Macfarlane, W. Meira, T. B. Schön, A.L.P. Ribeiro, Automatic diagnosis of the 12-lead ECG using a deep neural network, Nat. Commun. 11 (1) (Apr. 2020).

[9] N. Strodthoff, P. Wagner, T. Schaeffter, W. Samek, Deep learning for ECG analysis: benchmarks and insights from PTB-XL, IEEE J. Biomed. Health Inform. 25 (5) (2021) 1519–1528.

[10] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, M. Norouzi, Big Self-Supervised Models Advance Medical Image Classification, *arXiv preprint 2101/05224*, 2021.

[11] S. Hong, Y. Zhou, J. Shang, C. Xiao, J. Sun, Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review, Comput. Biol. Med. 122 (Jul. 2020), 103801.

[12] A. Mincholé, B. Rodriguez, Artificial intelligence for the electrocardiogram, Nat. Med. 25 (1) (Jan. 2019) 22–23.

[13] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F.I. Lunze, W. Samek, T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset, Sci. Data 7 (1) (2020) 154.

[14] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, A.Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nat. Med. 25 (1) (Jan. 2019) 65–69.

[15] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[16] X. Chen, H. Fan, R. Girshick, K. He, Improved Baselines with Momentum Contrastive Learning, 2020 arXiv preprint 2003.04297.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P.H. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z.D. Guo, M.G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent: a new approach to self-supervised learning, in: Advances in Neural Information Processing Systems, 2020.

[18] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, in: Advances in Neural Information Processing Systems, 2020.

[19] Y.-A. Chung, W.-N. Hsu, H. Tang, J. Glass, An unsupervised autoregressive model for speech representation learning, in: Proc. Interspeech 2019, 2019, pp. 146–150.

[20] M.A.C. Blandón, O. Räsänen, "Analysis of predictive coding models for phonemic representation learning in small datasets, in: ICML Workshop on Self-Supervision in Audio and Speech, 2020.

[21] L. van Staden, H. Kamper, A comparison of self-supervised speech representations as input features for unsupervised acoustic word embeddings, in: 2021 IEEE Spoken Language Technology Workshop (SLT), plus 0.5em minus 0.4emIEEE, 2021, pp. 927–934.

[22] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: unsupervised pre-training for speech recognition, in: Proc. Interspeech 2019, Apr 2019.

[23] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: a framework for self-supervised learning of speech representations, in: Advances in Neural Information Processing Systems, 2020.

[24] W. Falcon, K. Cho, A Framework for Contrastive Self-Supervised Learning and Designing a New Approach, *arXiv preprint arXiv:2009.00104*, 2020.

[25] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, M. Li, Bag of tricks for image classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 558–567.

[26] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[27] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Association for Computational Linguistics, Australia, Jul. 2018, pp. 328–339.

[28] Y. Yuan, G. Xun, Q. Suo, K. Jia, A. Zhang, Wave2vec: deep representation learning for clinical temporal data, Neurocomputing 324 (Jan. 2019) 31–42.

[29] J.Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, E. Azemi, Subject-aware Contrastive Learning for Biosignals, 2020 arXiv preprint 2007.04871.

[30] P. Sarkar, A. Etemad, Self-supervised ECG representation learning for emotion recognition, IEEE Trans. Affect. Comput. (2020) awaiting publication.

[31] D. Kiyasseh, T. Zhu, D.A. Clifton, Clocs: Contrastive Learning of Cardiac Signals across Space, Time, and Patients, 2020 arXiv preprint 2005.13249.

[32] H. Banville, I. Albuquerque, A. Hyvarinen, G. Moffat, D.-A. Engemann, A. Gramfort, "Self-supervised representation learning from electroencephalography signals, in: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, Oct. 2019.

[33] H. Banville, O. Chehab, A. Hyvarinen, D. Engemann, A. Gramfort, Uncovering the structure of clinical eeg signals with self-supervised learning, J. Neural. Eng. 18 (4) (2020), 046020.

[34] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi, Q. Li, A. Sharma, G. D. Clifford, and M. A. Reyna, Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020, Physiol. Meas. 41(12) (Jan. 2021) 124003.

[35] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, C. Rakovski, A 12-lead electrocardiogram database for arrhythmia research covering more than 10, 000 patients, Sci. Data 7 (1) (Feb. 2020).

[36] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J. E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet, Circulation 101 (23) (2000) e215–e220.

[37] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019.

[38] T. Mehari, N. Strodthoff [Online]. Available: https://github.com/hhi-aml/ecg-selfsupervised, 2021.

[39] G. Friesen, T. Jannett, M. Jadallah, S. Yates, S. Quint, H. Nagle, A comparison of the noise sensitivity of nine QRS detection algorithms, IEEE Trans. Biomed. Eng. 37 (1) (1990) 85–98.

[40] G. Lenis, N. Pilia, A. Loewe, W.H.W. Schulze, O. Dössel, Comparison of baseline wander removal techniques considering the preservation of ST changes in the ischemic ECG: a simulation study, Comput. Math. Methods Med. 2017 (2017) 1–13.